Numerical Comparison of Network Design Algorithms for Regionalized Variables

Jesus Carrera

Department of Hydrology and Water Resources
University of Arizona
Tucson, Arizona 85721

and

Ferenc Szidarovszky

Department of Mathematics and Computer Science
University of Horticulture
1118 Budapest, Villányi ut 29-35, Hungary

Transmitted by John Casti

ABSTRACT

The technique of kriging has a fundamental importance in applied sciences such as hydrology, meteorology, soil sciences, and mining. By using kriging, not only can the estimates of the natural phenomena be determined, but the estimation variances reflect the uncertainty of the estimation process. The sampling points for kriging should be selected to minimize the uncertainty, that is, to minimize the estimation variance of the kriging estimator. In this paper four algorithms for optimal observation network design are compared. Two of the algorithms give global optima, while the others give only suboptimal solutions. The computer time required for using the heuristic algorithms, giving only suboptimal solutions, is much less than that of the optimizing procedures. It is also shown on the basis of our experiments that the suboptimal solutions are either optimal or very close to optimal; consequently on the basis of our simulated examples, the heuristic algorithms are highly recommended for practical applications.

INTRODUCTION

In applied sciences a great deal of data has to be collected and analyzed. In many applications these data are very expensive, e.g. drillhole data, which are

52 Vanderbilt Ave., New York, NY 10017 0096-3003/85/$03.30

used in mining exploration and other geosciences. In these applications the optimal location of measurement points is a very important problem because of the large expense of collecting data.

The optimal observation network design procedures are mostly based on the theory of regionalized variables, which has been developed by Matheron [5–8]. On the basis of his theory a number of estimation processes have been introduced, which are called the different variants of kriging. In this paper the application of the "classical" kriging method will be examined; the more sophisticated variants, such as universal kriging and cokriging, can be investigated in the same manner.

In applications it is usually assumed that $Z(x)$ is a random variable for all values of $x \in D$, where D is the domain of the stochastic function $Z(\cdot)$. A function $Z(\cdot)$ is called *intrinsic* if

$$E\{Z(x+h)-Z(x)\}=0, \tag{1}$$

$$\mathrm{Var}\{Z(x+h)-Z(x)\}=2\gamma(h) \tag{2}$$

for all $x, x+h \in D$. In these hypotheses the increments of the function, $Z(x+h)-Z(x)$, rather than the function itself are considered. The first assumption does not imply the existence of the expected value of $Z(x)$, since in the case when $Z(x)$ has the same Cauchy distribution for all x, $Z(x+h)-Z(x)=0$ for all x and h, but $E\{Z(x)\}$ does not exist. In the second assumption it is assumed that the variance of the increment $Z(x+h)-Z(x)$ depends only on h. The function $\gamma(\cdot)$ is called the *variogram*. The main properties and the usual forms of the variogram function are discussed in [4].

Let V denote a subset of D, which is called for example a *block* in the mining industry. The average value of the function $Z(\cdot)$ on the block V is given as

$$\bar{Z}(V)=\frac{1}{|V|}\int_V Z(x)\,dx, \tag{3}$$

where $|V|$ denotes the length, area, or volume of V in the one, two, or three dimensional case, respectively. In kriging this average value is estimated by the linear form

$$Z^*=\sum_{j=1}^{n}\lambda_j Z(x_j), \tag{4}$$

where $x_1,\ldots,x_n$ are distinct measurement points on D and $Z(x_1),\ldots,Z(x_n)$

are the corresponding measurement values. It is required that the estimator (4) be unbiased and optimal (i.e. with minimal squared error). The unbiasedness condition is equivalent to the relation

$$\sum_{j=1}^{n} \lambda_j = 1, \tag{5}$$

and the mean squared error is given as

$$\begin{aligned} \mathrm{Var}\{Z^*\} &= E\{Z^* - \bar{Z}(V)\} \\ &= -\gamma_{VV} + 2\sum_{i=1}^{n} \lambda_i \gamma_{Vi} - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma_{ij}, \end{aligned} \tag{6}$$

where

$$\gamma_{ij} = \gamma(x_i - x_j),$$

$$\gamma_{Vi} = \frac{1}{|V|}\int_V \gamma(x - x_i)\,dx,$$

$$\gamma_{VV} = \frac{1}{|V|^2}\int_V\int_V \gamma(x - x')\,dx\,dx'.$$

It is well known from the kriging literature (e.g. [4, 2]) that the minimization of the estimation variance (6) subject to the contraint (5) is equivalent to the solution of the linear system of equations

$$\sum_{j=1}^{n} \lambda_j \gamma_{ij} + \mu = \gamma_{Vi} \qquad (i = 1,2,\ldots,n),$$

$$\sum_{j=1}^{n} \lambda_j = 1. \tag{7}$$

which are called the *kriging equations*. If $\lambda_1^*,\ldots,\lambda_n^*, \mu^*$ denote the solution of the equations (7), then the corresponding estimation variance has the

simple linear form

$$\mathrm{Var}\{Z^*\} = \mu + \sum_{j=1}^{n} \lambda_j^* \gamma_{Vj} - \gamma_{VV}, \tag{8}$$

which is called the *estimation variance.*

The particular form of the kriging equations and the estimation variance imply the following observations:

(1) The value of the estimation variance for a given variogram depends only on the selection of the block V and the measurement locations $x_1, \ldots, x_n$, and does not depend on the measurements $Z(x_1), \ldots, Z(x_n)$. Thus, the estimation variance can be computed before the actual measuring process is performed.

(2) Assume that a new observation point x_{n+1} is included into the set of measurement points. Then

(a) the estimation variance decreases;

(b) it can be updated without repeating the entire calculations by using the method known as "inversion by blocks" (see [9, 1]).

(3) Assume next that a point is dropped from the set of measurement locations. Then

(a) the estimation variance increases;

(b) it can be updated by a procedure similar to that used in the previous case.

On the basis of these properties of kriging, four observation network procedures will be introduced in the next section.

OBSERVATION NETWORK DESIGN PROCEDURES

In this section four algorithms will be introduced, namely

(1) total enumeration,
(2) branch and bound algorithm,
(3) sequential optimal including of additional points,
(4) sequential optimal exchanges.

Let $x_1, \ldots, x_k$ denote the existing measurement points, and assume that $n - k$ further measurement points are to be selected. Let $t_1, \ldots, t_N$ denote the candidates for the additional points, and assume that $N > n - k$. The optimal selection of additional points from the finite set $T = \{t_1, \ldots, t_N\}$ can be mathematically formulated in the following way: Find the elements $t_{i_1}, \ldots, t_{i_{n-k}}$

from T such that the estimation variance based on the measurement points

$$x_1, \ldots, x_k, t_{i_1}, \ldots, t_{i_{n-k}}$$

is minimal.

In applying *total enumeration*, all subsets of T having $n-k$ elements are systematically generated. For each subset, the corresponding estimation variance is computed and the subset giving the smallest value of the estimation variance is accepted as the optimal solution. Observe that the number of subsets to be searched equals $\binom{N}{n-k}$, and a systematic search procedure can be performed, e.g., in the following way. The nodes of the search tree correspond to the subsets having at most $n-k$ elements of T. The initial node corresponds to the empty set (i.e., when no additional point is included in the kriging procedure), and a directed arc connects the subsets T_1, T_2 of T if and only if $T_2 - T_1$ has only one element which has higher subscript than the subscripts of all elements of T_1. Figure 1 illustrates the search tree for $N=5$, $n-k=3$. The search procedure starts at the initial node (which corresponds to the empty set $\varnothing$), and finding ourselves at any node, we check whether at least one of the following conditions holds:

(i) the number of the elements of the subset of T which corresponds to this node equals $n-k$;

(ii) all of the nodes which are endpoints of arcs starting from the current node have already been searched.

If either condition (i) or condition (ii) holds, then we should proceed backwards; otherwise we should proceed forward to the next point which has not been searched so far during the procedure. Observe that moving forward along any arc is equivalent to adding one point to the kriging process, and moving backwards along any arc is equivalent to dropping the point having the largest subscript. The estimation variance can be updated in both cases by using the numerical procedure mentioned at the end of the previous section. Observe furthermore that the subsets of T having exactly $n-k$ elements are among the endpoints of the search tree. Hence the corresponding updated estimation variances should be compared, and the smallest of them should be chosen to select the optimal solution.

The idea of the *branch and bound* procedure is very similar to that of total enumeration, but there are two major differences, which are now discussed. The first difference is given by the construction of the search tree. In this case the initial node corresponds to the entire set T, and moving along each arc is equivalent to dropping one point from the kriging process. The second difference is given by the conditions to be checked for deciding whether we

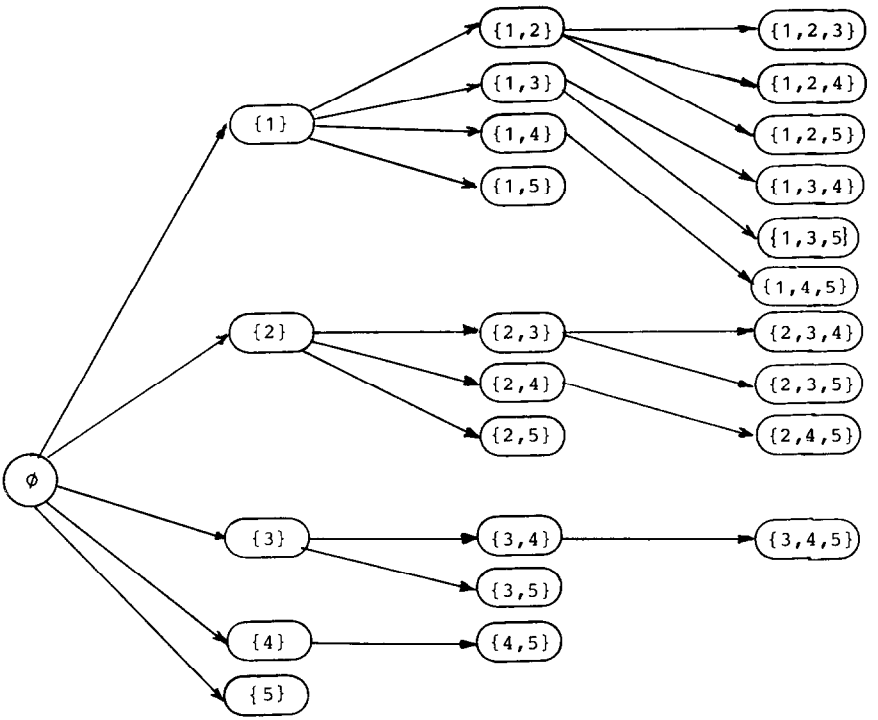

FIG. 1. Illustration of the search tree.

should move backwards or not. In this case an additional condition should be checked:

(iii) the current estimation variance is not less than the smallest one found for subsets having exactly $n - k$ element.

In this case dropping additional points makes the estimation variance even larger.

If at least one of conditions (i), (ii), and (iii) holds, then we should proceed backwards; otherwise we should proceed forward, as in total enumeration [1].

The *method of sequential including* starts from the empty set $\emptyset$, that is, it starts from the case when no additional measurement point is included in the kriging. In other words, first only the measurement points $x_1, \ldots, x_k$ are considered. Then, for $j = 1, \ldots, N$, the sets $\{x_1, \ldots, x_k, t_j\}$ are examined, and the point t_{j_1} which gives the smallest estimation variance is selected as the first additional observation point. After the point t_{j_1} has been selected, all of the subsets $\{x_1, \ldots, x_k, t_{j_1}, t_j\}$ $(j \neq j_1, 1 \leqslant j \leqslant N)$ are examined, and the point

t_{j_2} which gives the smallest estimation variance is selected as the second additional measurement point. After determining t_{j_2}, all the sets $\{x_1, \ldots, x_k, t_{j_1}, t_{j_2}, t_j\}$ $(j \neq j_1,\ j \neq j_2,\ 1 \leqq j \leqq N)$ are examined, and so on. This procedure is continued until exactly $n-k$ new points have been included. In including any one of the additional measurement points the estimation variance can be updated by using the simplified procedure mentioned at the end of the previous section.

The *method of sequential optimal exchanges* is based on the following principle. Let $X_0 = \{t_1, \ldots, t_{n-k}\}$, $X_1 = \{t_{n-k+1}, \ldots, t_N\}$, and $j = 1$. Then try to exchange the jth element of X_0 systematically with the elements of X_1; the exchange that minimizes the estimation variance will be actually performed. If this optimal exchange has been performed, then let X_0 denote the new set, and let X_1 denote the set of the remaining points. If no exchange can decrease the estimation variance, then do not modify sets X_0 and X_1. But in both cases modify the value of j by the following rule:

$$j := \begin{cases} j+1 & \text{if} \quad j < n-k, \\ 1 & \text{if} \quad j = n-k, \end{cases}$$

and try to exchange the jth element of X_0 optimally, and so on. The procedure terminates if no element of X_0 can be exchanged to decrease the estimation variance.

Note that the first two algorithms give the optimal selection of the additional measurement points, and the last two algorithms do not necessarily give an optimal solution. On the other hand, the last two methods usually need much less computer time than either of the first two. Hence, the selection of the method to be applied in practical problems should be based on the decision maker's subjective judgment in balancing computer time against the acceptance of suboptimal solutions. Such a decision should be based on computational experience in using all of the above-discussed methods in practical problems, with comparison of the results and execution times. In the next part of this paper such computational experiments will be reported.

NUMERICAL EXAMPLES

The methodology described in the previous section was applied to the estimation of the average log transmissivity in a region of the Tajo Basin (Spain), west of Madrid. Figure 2 shows the relative location of the existing wells for which data are available and for which no data are available, as well as the block V over which the average log transmissivity is to be estimated [3].

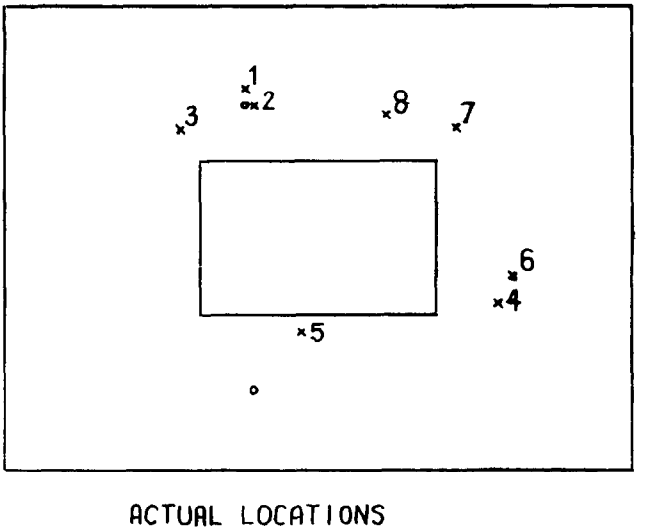

FIG. 2. Location of existing wells. Circles represent wells with available data. The interior rectangle is the region over which $\log T$ is to be estimated.

The block V given by the rectangle having corners (57.5, 22.5), (57.5, 32.5), (72.5, 22.5), and (72.5, 32.5) is examined. Two existing measurement points are given by the locations of two wells, where actual measurements are performed. Thus, $k = 2$, $x_1 = (60.6, 36.2)$, $x_2 = (60.9, 17.7)$. Eight possibilities for additional measurement points are given by the further existing wells in the near neighborhood of the subregion V. Hence $N = 8$ and

$$t_1 = (60.4, 37.2), \quad t_2 = (60.9, 36.1), \quad t_3 = (56.3, 34.6),$$

$$t_4 = (76.4, 23.3), \quad t_5 = (63.9, 21.5), \quad t_6 = (77.4, 25.3),$$

$$t_7 = (73.9, 34.8), \quad t_8 = (69.4, 35.6).$$

The best four points from this finite set are selected by using all of the four methods described in the previous section. The variogram is selected as [3]

$$\gamma(h) = \begin{cases} 0 & \text{if } |h| = 0, \\ 0.08 + 0.1\left[\dfrac{3}{2}\dfrac{|h|}{40} - \dfrac{1}{2}\left(\dfrac{|h|}{40}\right)^3\right] & \text{if } 0 < |h| < 40. \\ 0.18 & \text{if } |h| \geqslant 40. \end{cases}$$

TABLE 1

ADDITIONAL MEASUREMENT POINT CANDIDATES IN 10 SIMULATION CASES

Example 1		Example 6	
55.08	39.21	65.73	28.42
69.75	21.85	60.64	27.25
56.65	39.53	76.16	37.40
63.12	39.02	76.93	34.41
68.29	26.52	56.20	29.53
65.21	32.47	59.52	34.31
62.23	32.71	59.77	27.73
72.79	24.92	62.47	33.17
Example 2		Example 7	
77.06	21.71	77.30	33.95
69.91	39.21	65.73	20.26
60.10	38.31	74.79	33.56
55.55	28.39	66.98	22.01
71.95	33.19	56.63	39.79
70.47	36.19	74.74	33.28
68.73	26.17	73.51	29.41
57.51	22.75	62.63	32.18
Example 3		Example 8	
71.42	21.18	74.96	30.00
74.09	38.85	77.90	20.92
71.56	36.20	65.12	22.04
73.24	23.62	76.64	35.12
74.34	35.97	70.06	26.36
60.96	20.52	55.93	29.02
69.91	33.91	59.39	32.81
77.12	31.62	67.86	26.02
Example 4		Example 9	
63.47	24.88	66.04	26.49
59.77	33.92	77.38	30.92
58.04	28.12	57.00	25.50
66.99	28.13	59.38	31.76
79.51	30.62	77.65	23.70
66.82	20.86	63.36	30.96
56.56	38.47	62.32	22.10
64.87	27.04	72.07	24.52
Example 5		Example 10	
74.51	25.56	72.15	26.16
64.48	29.41	58.86	37.31
79.01	36.19	76.18	39.33
65.98	25.18	69.22	36.92
65.32	36.19	76.62	36.50
77.61	21.74	77.59	38.12
65.70	24.73	79.69	36.98
71.08	22.09	59.98	37.59

The computations have been repeated in ten more cases, where only the eight candidates for the additional measurement points were exchanged. They were generated randomly in the neighborhood of V. The actual point coordinates are given in Table 1 and plotted in Figure 3.

The optimal selection of addition observation points, minimal estimation variances, and execution times are summarized in Table 2. From these results, the following conclusions can be drawn.

(1) The execution time of total enumeration depends only slightly on the locations of the existing observation points and the candidates for additional points. The number of operations grows exponentially with the number of the candidates for additional measurement points, which was denoted by N.

(2) The execution time of the branch and bound method significantly depends on the locations of the points. The number of operations grows at most exponentially if N increases, but in special cases it may grow much less fast.

(3) The previous observation also holds for the method of sequential exchanges, since the maximal number of exchanges equals $\binom{N}{n-k}-1$. But in special cases, e.g. if one starts from the optimal solution, then many less operations are needed. In this case, after performing one cycle, the algorithm terminates. Observe that in our examples the execution time was always small, comparable to sequential including.

(4) In the case of sequential including the number of operations is only polynomial. In order to verify this assertion, observe that the number of estimation variances which are computed and compared equals

$$N+(N-1)+\cdots+(N-n+k-1)=\frac{2N-n+k-1}{2}\cdot(n-k),$$

updating one estimation variance of a kriging estimate based on $k+l$ $(l=1,\ldots,n-k)$ points requires $O(l^2)$ operations, and the determination of the initial estimation variance requires $O(k^3)$ operations.

(5) In our case $k=n-k$; consequently the branch and bound tree and the search tree for total enumeration have exactly the same size. Both methods give the optimum, but the execution time of total enumeration was at least 10% more than that of branch and bound algorithm. The highest relative execution time difference was about 47% (in the case of the third simulation example). On the basis of our result the following suggestion can be made. If the branch and bound tree is not substantially larger than the search tree for total enumeration, then the branch and bound method should be selected.

(6) Observe that the cheaper branch and bound method needs about 6–10 times more computer time than either of the last two methods. Note that

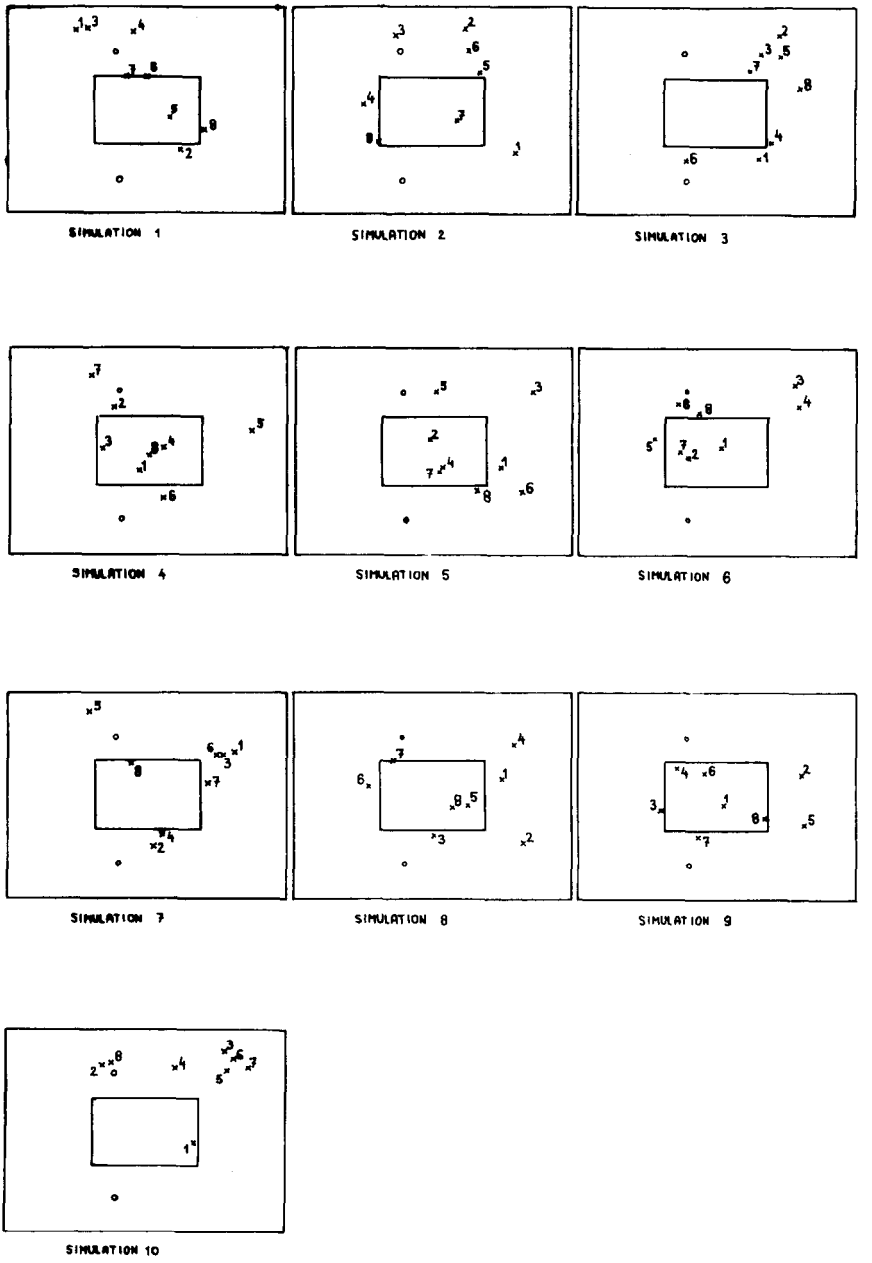

FIG. 3. Display of the 10 simulated sets of eight wells. In each plot, circles represent locations of wells with available data.

TABLE 2

SUMMARY OF RESULTS AND EXECUTION TIMES

	Branch and bound			Total enumeration			Sequential including			Sequential exchanges		
	Optimal selection	Optimal variance	Execution time (sec)	Optimal selection	Optimal variance	Execution time (sec)	Optimal selection	Optimal variance	Execution time (sec)	Optimal selection	Optimal variance	Execution time (sec)
Real data	3, 4, 5, 8	0.0319	0.3680	3, 4, 5, 8	0.0319	0.5100	3, 5, 6, 8	0.0320	0.0520	3, 4, 5, 8	0.0319	0.0650
Simulation:												
No. 1	5, 6, 7, 8	0.0287	0.3600	5, 6, 7, 8	0.0287	0.5010	5, 6, 7, 8	0.0287	0.0450	5, 6, 7, 8	0.0287	0.0580
No. 2	4, 5, 7, 8	0.0294	0.4030	4, 5, 7, 8	0.0294	0.5050	4, 5, 7, 8	0.0294	0.0450	4, 5, 7, 8	0.0294	0.0550
No. 3	1, 4, 6, 7	0.0315	0.3400	1, 4, 6, 7	0.0315	0.5020	1, 4, 6, 7	0.0315	0.0450	1, 4, 6, 7	0.0315	0.0600
No. 4	3, 4, 5, 8	0.0282	0.4140	3, 4, 5, 8	0.0282	0.5010	3, 4, 5, 8	0.0282	0.0450	3, 4, 5, 8	0.0282	0.0600
No. 5	1, 2, 4, 7	0.0283	0.4180	1, 2, 4, 7	0.0283	0.5010	1, 2, 4, 7	0.0283	0.0480	1, 2, 4, 7	0.0283	0.0550
No. 6	1, 2, 4, 7	0.0292	0.4370	1, 2, 4, 7	0.0292	0.5000	1, 2, 4, 7	0.0292	0.0450	1, 2, 4, 7	0.0292	0.0520
No. 7	2, 4, 7, 8	0.0296	0.3970	2, 4, 7, 8	0.0296	0.5030	2, 4, 7, 8	0.0296	0.0450	2, 4, 7, 8	0.0296	0.0530
No. 8	1, 5, 6, 8	0.0284	0.4450	1, 5, 6, 8	0.0284	0.4980	1, 5, 6, 8	0.0284	0.0460	1, 5, 6, 8	0.0284	0.0580
No. 9	1, 3, 6, 8	0.0277	0.4550	1, 3, 6, 8	0.0277	0.4990	1, 3, 6, 8	0.0277	0.0450	1, 3, 6, 8	0.0277	0.0630
No. 10	1, 2, 4, 5	0.0350	0.4120	1, 2, 4, 5	0.0350	0.5010	1, 2, 4, 5	0.0350	0.0460	1, 2, 4, 5	0.0350	0.0550

sequential including gives the optimum in almost all of the cases, and sequential exchange gives the optimum in each of the cases. On the basis of these observation we can recommend the use of any of these methods if the decision maker is satisfied with a reasonably good suboptimal solution.

(7) Furthermore, even in the case when the decision maker requires the exact optimum, the computations should be performed in two stages: first one of the inexpensive methods should be applied, and its solution should be selected as the top line of the branch and bound tree. This selection will increase the frequency of cases when the estimation variances of earlier stages are worse than the best one found so far, since in the first line a very good (closely optimal) estimation variance has been found.

(8) In applying nonoptimal methods, a two stage procedure is also suggested. First apply sequential including, and select its solution for the initial set of additional observation points as starting points for sequential exchanges. This combination about doubles the execution time, but the doubled execution time will still be a lot less than the execution time needed by any of the optimal methods.

CONCLUSIONS

Four methods for the design of measurement networks have been presented and their merits compared. From such comparison, the following conclusions can be drawn.

(1) The total enumeration and branch and bound methods lead to optimal solutions. The latter will be more efficient than the former unless its search tree is much larger than that of the total enumeration.

(2) Sequential including and sequential exchange give almost optimal solution at a cost that is an order of magnitude smaller.

(3) The cost of the branch and bound method depends on the choice of the first sets searched. A preliminary run with one of the suboptimal methods is proposed in order to improve the choice of the first set to search and to reduce the estimation variance of the first stages of the branch and bound tree.

(4) A very efficient, although not optimal, design can be obtained by selecting the result obtained by sequential including as the first set X_0 for sequential exchange.

REFERENCES

1 J. Carrera, E. Usunoff, and F. Szidarovszky, Optimal observation network design for ground water management application to the San Pedro River Basin, Arizona,

submitted for publication.
2 M. David, *Geostatistical Ore Reserve Estimation*, Elsevier, Amsterdam, 1977.
3 P. J. Fennessy, Geostatistical analysis and stochastic modeling of the Tajo Basin aquifer, Spain, unpublished M.S. Thesis, Dept. of Hydrology and Water Resources, Univ. of Arizona, Tucson, 1982.
4 A. G. Journel and Ch. J. Huijbregts, *Mining Geostatistics*, Academic, New York, 1978.
5 G. Matheron, Théorie lognormale de l'échantillonnage systématique des gisements, *Ann. Mines*, 1957.
6 G. Matheron, *Traité de Géostatistique Appliquée*, Vols. 1, 2, Editions Technio, Paris, 1963.
7 G. Matheron, Le krigeage universel, Cahiers du Centre de Morphologie Math., No. 1, ENSMP, Paris, 1969.
8 G. Matheron, The theory of regionalized variables and its applications, Cahiers du Centre de Morphologie Math., No. 5, ENSMP, Paris, 1971.
9 F. Szidarovszky and S. Yakowitz, *Principles and Procedures of Numerical Analysis*, Plenum, New York, 1978.